Matching arrays using the pathway fingerprint

Gabriel Altschuler

June 2, 2013

In this document we will test the ability of the pathway fingerprint to pull out simliar arrays across platforms and species in the GEO corpus.
The required data and metadata is contained within the R package `pathprint`. In addition, we will make use of the tissue samples curated in *Zilliox and Irizarry. A gene expression bar code for microarray data. Nat Meth (2007) vol. 4 (11) pp. 911-3*, a set of arrays from 6 tissues; muscle, lung, spleen, kidney, liver and brain.
First we need to source the pathprint package and load the data libraries containing the fingerprint and the metadata. We will also load the tissue-specific data from a local file.

```
> library(pathprint)
> tissue.meta<-read.delim(
+   "barcode_figure2_data.txt",
+   stringsAsFactors = FALSE)
> colnames(tissue.meta)
[1] "X"             "filename"      "DB_ID"         "ExperimentID"
[5] "Tissue"        "SubType"       "ClinicalGroup" "Platform"
> tissue.meta<-tissue.meta[
+   tissue.meta$DB_ID %in% colnames(GEO.fingerprint.matrix),]
> (tissues<-levels(as.factor(tissue.meta$Tissue)))
[1] "brain"           "kidney"          "liver"           "lung"
[5] "skeletal muscle" "spleen"
```

1 Tissue consensus fingerprints

For each tissue, we will construct a "*consensus fingerprint*", this is a pathway vector that contains `1` or `-1` at for pathways that have consistently high or low expression across all of the samples in the set and `0` otherwise, given a fractional threshold. This uses the pathprint functions `consensusFingerprint`, We will initailly set the threshold to 0.75.

```
> tissueConsensus<-sapply(tissues, function(x){
+     consensusFingerprint(GEO.fingerprint.matrix[,
+         tissue.meta$DB_ID[tissue.meta$Tissue == x]],
+     threshold = 0.75)
+     })
> head(tissueConsensus)
```

```
                                                brain kidney liver lung
Glycolysis / Gluconeogenesis (KEGG)                 0      1     1    0
Citrate cycle (TCA cycle) (KEGG)                    0      1     1   -1
Pentose phosphate pathway (KEGG)                   -1      1     1    0
Pentose and glucuronate interconversions (KEGG)    -1      1     1    1
Fructose and mannose metabolism (KEGG)              0      1     0    0
Galactose metabolism (KEGG)                         0      0     1    0
                                                skeletal muscle spleen
Glycolysis / Gluconeogenesis (KEGG)                            1     -1
Citrate cycle (TCA cycle) (KEGG)                               1      0
Pentose phosphate pathway (KEGG)                              -1      0
Pentose and glucuronate interconversions (KEGG)                0      0
Fructose and mannose metabolism (KEGG)                         0     -1
Galactose metabolism (KEGG)                                    0      0
```

We can now use the Manhattan distance as a measure of similiarity to each of these vectors for all of the arrays in the fingerprint corpus. The function `consensusDistance` is used to calculate the distances and approximate p-value.

```
> tissueDistance<-apply(tissueConsensus, 2,
+         consensusDistance, GEO.fingerprint.matrix)
[1] "Scaling against max length, 514"
[1] "Scaling against max length, 308"
[1] "Scaling against max length, 526"
[1] "Scaling against max length, 310"
[1] "Scaling against max length, 330"
[1] "Scaling against max length, 232"
```

An alternative method is to use the mean fingerprint and calculate the distance from this The status outputs indicate the maximum manhattan distance possible based on the number of scored pathways in each consensus.

```
> op<-par(mfrow = c(1,1), pty = "s")
> par(mfcol = c(2,3))
> brain.dist.hist<-hist(tissueDistance$brain[,"distance"],
+     nclass = 50, xlim = c(0,1), main = "brain", xlab = "Scaled distance")
> kidney.dist.hist<-hist(tissueDistance$kidney[,"distance"],
+     nclass = 50, xlim = c(0,1), main = "kidney", xlab = "Scaled distance")
> liver.dist.hist<-hist(tissueDistance$liver[,"distance"],
+     nclass = 50, xlim = c(0,1), main = "liver", xlab = "Scaled distance")
> lung.dist.hist<-hist(tissueDistance$lung[,"distance"],
+     nclass = 50, xlim = c(0,1), main = "lung", xlab = "Scaled distance")
> skeletalmuscle.dist.hist<-hist(tissueDistance$"skeletal muscle"[,"distance"],
+     nclass = 50, xlim = c(0,1), main = "skeletal muscle", xlab = "Scaled distance")
> spleen.dist.hist<-hist(tissueDistance$spleen[,"distance"],
+     nclass = 50, xlim = c(0,1), main = "spleen", xlab = "Scaled distance")
> par(op)


> hist(tissueDistance$brain[
+         tissue.meta$DB_ID[tissue.meta$Tissue == "brain"],
+         "distance"],
+     breaks = brain.dist.hist$breaks,
+     xlim = c(0,1), main = "", xlab = "Scaled distance")
> #par(op)
```

Figure 2 shows a histogram of the distances from the brain consensus for the full corpus and for the brain tissues that made up the signature. This can also be represented as a table. Or we can extract the top ranking arrays for a particular platform, e.g. what are the closest matched *Danio Rerio*, *Rattus norvegicus*, *Drosophila melanogaster*, *Caenorhabditis elegans* arrays to the human/mouse brain and liver consensus?

```
> brain.meta<-cbind(tissueDistance$brain,
+               GEO.metadata.matrix[
+                 match(rownames(tissueDistance$brain),
+                 GEO.metadata.matrix$GSM)
+               ,])
> Dr.brain.match<-head(brain.meta[
+                       brain.meta$Species == "Danio rerio",
+                       c(1,2,5,6,8)
+                     ],20)


> liver.meta<-cbind(tissueDistance$liver,
+               GEO.metadata.matrix[
+                 match(rownames(tissueDistance$liver),
+                 GEO.metadata.matrix$GSM)
+               ,])
> Dr.liver.match<-head(liver.meta[
+                      liver.meta$Species == "Danio rerio",
+                      c(1,2,5,6,8)
+                    ],20)


> Rn.brain.match<-head(brain.meta[
+                      brain.meta$Species == "Rattus norvegicus",
+                      c(1,2,5,6,8)
+                    ],20)


> Rn.liver.match<-head(liver.meta[
+                      liver.meta$Species == "Rattus norvegicus",
+                      c(1,2,5,6,8)
+                    ],20)


> Ce.brain.match<-head(brain.meta[
+                       brain.meta$Species == "Caenorhabditis elegans",
+                       c(1,2,5,6,8)
+                     ],20)
```

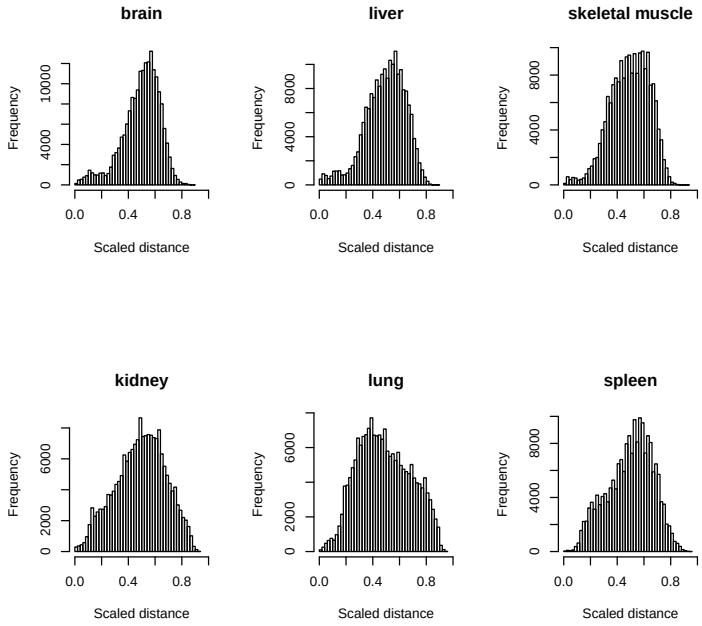

Figure 1: Histograms of the distance from the tissue consensus of all GEO samples

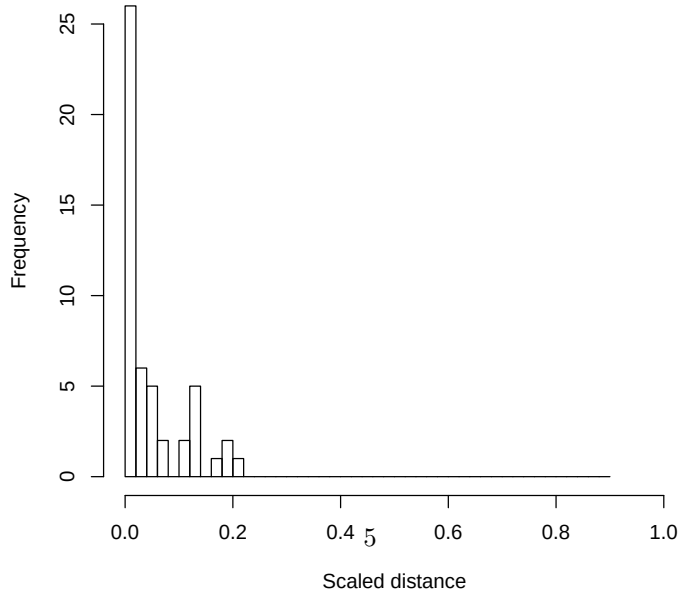

Figure 2: Histograms of the distance from the brain consensus of the brain tissues that made up the consensus signature.

```
> Ce.liver.match<-head(liver.meta[
+                      liver.meta$Species == "Caenorhabditis elegans",
+                      c(1,2,5,6,8)
+                    ],20)


> Dm.brain.match<-head(brain.meta[
+                        brain.meta$Species == "Drosophila melanogaster",
+                        c(1,2,5,6,8)
+                      ],20)


> Dm.liver.match<-head(liver.meta[
+                      liver.meta$Species == "Drosophila melanogaster",
+                      c(1,2,5,6,8)
+                    ],20)
```

Table 1: Human/mouse brain - zebrafish matches

	distance	pvalue	GPL	Species	Source
GSM337585	0.224	0.025	GPL1319	Danio rerio	3d old, brain, night-time
GSM337581	0.232	0.029	GPL1319	Danio rerio	3d old, pineal gland, night-time
GSM337584	0.232	0.029	GPL1319	Danio rerio	3d old, brain, night-time
GSM305894	0.233	0.030	GPL1319	Danio rerio	brain
GSM337583	0.233	0.030	GPL1319	Danio rerio	3d old, brain, night-time
GSM305896	0.235	0.031	GPL1319	Danio rerio	brain
GSM305898	0.235	0.031	GPL1319	Danio rerio	brain
GSM337575	0.235	0.031	GPL1319	Danio rerio	3d old, brain, day-time
GSM337592	0.235	0.031	GPL1319	Danio rerio	5d old, brain, day-time
GSM305893	0.237	0.032	GPL1319	Danio rerio	brain
GSM305897	0.237	0.032	GPL1319	Danio rerio	brain
GSM337593	0.237	0.032	GPL1319	Danio rerio	5d old, brain, day-time
GSM337598	0.237	0.032	GPL1319	Danio rerio	5d old, brain, night-time
GSM305899	0.239	0.033	GPL1319	Danio rerio	brain
GSM337582	0.241	0.034	GPL1319	Danio rerio	3d old, brain, night-time
GSM305891	0.243	0.035	GPL1319	Danio rerio	brain
GSM337596	0.243	0.035	GPL1319	Danio rerio	5d old, pineal gland, night-time
GSM305892	0.245	0.036	GPL1319	Danio rerio	brain
GSM337633	0.247	0.037	GPL1319	Danio rerio	1-2yr old, brain, day-time
GSM305895	0.251	0.039	GPL1319	Danio rerio	brain

Table 2: Human/mouse brain - rat matches

	distance	pvalue	GPL	Species	Source
GSM510570	0.078	0.001	GPL1355	Rattus norvegicus	Frontal cortex from Flinders Depression Sensitive Line (FRL) rats
GSM136499	0.080	0.001	GPL1355	Rattus norvegicus	amygdala
GSM136464	0.082	0.001	GPL1355	Rattus norvegicus	amygdala
GSM136489	0.082	0.001	GPL1355	Rattus norvegicus	amygdala
GSM341339	0.082	0.001	GPL1355	Rattus norvegicus	Rat forebrain, control condition
GSM136460	0.084	0.002	GPL1355	Rattus norvegicus	hippocampus
GSM136469	0.086	0.002	GPL1355	Rattus norvegicus	amygdala
GSM136474	0.086	0.002	GPL1355	Rattus norvegicus	amygdala
GSM510567	0.086	0.002	GPL1355	Rattus norvegicus	Frontal cortex from Flinders Depression Sensitive Line (FRL) rats
GSM136454	0.088	0.002	GPL1355	Rattus norvegicus	amygdala
GSM136497	0.088	0.002	GPL1355	Rattus norvegicus	striatum
GSM136504	0.088	0.002	GPL1355	Rattus norvegicus	amygdala
GSM341334	0.088	0.002	GPL1355	Rattus norvegicus	Rat forebrain, control condition
GSM510550	0.088	0.002	GPL1355	Rattus norvegicus	Frontal cortex from Flinders Depression Resistent Line (FRL) rats
GSM510589	0.088	0.002	GPL1355	Rattus norvegicus	Hippocampus from Flinders Depression Resistent Line (FRL) rats
GSM136449	0.089	0.002	GPL1355	Rattus norvegicus	amygdala
GSM136467	0.089	0.002	GPL1355	Rattus norvegicus	striatum
GSM136479	0.089	0.002	GPL1355	Rattus norvegicus	amygdala
GSM136485	0.089	0.002	GPL1355	Rattus norvegicus	hippocampus
GSM136488	0.089	0.002	GPL1355	Rattus norvegicus	frontal_cortex

Table 3: Human/mouse brain - C. elegans matches

	distance	pvalue	GPL	Species	Source
GSM272407	0.214	0.022	GPL200	Caenorhabditis elegans	10_E5_C.Elegans microwave exposed
GSM589334	0.214	0.022	GPL200	Caenorhabditis elegans	C. elegans adults 30 hr time-point during DD free-running
GSM589373	0.214	0.022	GPL200	Caenorhabditis elegans	C. elegans adults 22 hr time-point during WC entrainment
GSM627501	0.214	0.022	GPL200	Caenorhabditis elegans	3-fold stage embryos
GSM589376	0.216	0.022	GPL200	Caenorhabditis elegans	C. elegans adults 10 hr time-point during WC entrainment
GSM589379	0.218	0.023	GPL200	Caenorhabditis elegans	C. elegans adults 22 hr time-point during WC entrainment
GSM197728	0.222	0.025	GPL200	Caenorhabditis elegans	RNA obtained from all neurons after co-IP from F25B3.3::FLAG::PAB-1 transge
GSM272403	0.224	0.025	GPL200	Caenorhabditis elegans	6_E3_C.Elegans microwave exposed
GSM589375	0.226	0.026	GPL200	Caenorhabditis elegans	C. elegans adults 6 hr time-point during WC entrainment
GSM589357	0.228	0.027	GPL200	Caenorhabditis elegans	C. elegans adults 2 hr time-point during non-entrained DD
GSM589391	0.232	0.029	GPL200	Caenorhabditis elegans	C. elegans adults 22 hr time-point during WC entrainment
GSM279203	0.237	0.032	GPL200	Caenorhabditis elegans	C. elegans 3-fold embryo
GSM589347	0.247	0.037	GPL200	Caenorhabditis elegans	C. elegans adults 34 hr time-point during DD free-running
GSM589393	0.249	0.038	GPL200	Caenorhabditis elegans	C. elegans adults 30 hr time-point during CC free-running
GSM197731	0.251	0.039	GPL200	Caenorhabditis elegans	RNA obtained from A-class neurons after co-IP from unc-4::3XFLAG::PAB-1 tra
GSM197734	0.253	0.040	GPL200	Caenorhabditis elegans	RNA obtained from A-class neurons after co-IP from unc-4::3XFLAG::PAB-1 tra
GSM453400	0.253	0.040	GPL200	Caenorhabditis elegans	Wild Type L1 animal
GSM453406	0.257	0.043	GPL200	Caenorhabditis elegans	oga mutant L1 animal
GSM453399	0.259	0.044	GPL200	Caenorhabditis elegans	Wild Type L1 animal
GSM453401	0.259	0.044	GPL200	Caenorhabditis elegans	ogt mutant L1 animal

Table 4: Human/mouse brain - Drosophila matches

	distance	pvalue	GPL	Species	Source
GSM398835	0.243	0.035	GPL1322	Drosophila melanogaster	DmWT CH # 3, normobaric hypoxia, two weeks
GSM398840	0.245	0.036	GPL1322	Drosophila melanogaster	DmWT Nx # 3, normoxia, two weeks
GSM368567	0.253	0.040	GPL1322	Drosophila melanogaster	D.m. adult heads yw
GSM368569	0.255	0.042	GPL1322	Drosophila melanogaster	D.m. adult heads yw
GSM398839	0.255	0.042	GPL1322	Drosophila melanogaster	DmWT Nx # 2, normoxia, two weeks
GSM368579	0.259	0.044	GPL1322	Drosophila melanogaster	D.m. adult heads egg
GSM80134	0.259	0.044	GPL72	Drosophila melanogaster	adult head
GSM101542	0.261	0.046	GPL72	Drosophila melanogaster	Heads of Drosophila received humidified air for 50 min followed by a single exp
GSM266585	0.261	0.046	GPL1322	Drosophila melanogaster	Male fly heads, WT control
GSM101547	0.265	0.048	GPL72	Drosophila melanogaster	Heads of Drosophila received humidified air for 50 min followed by 5 exposures
GSM296990	0.265	0.048	GPL1322	Drosophila melanogaster	Diet:Untreated
GSM440094	0.267	0.050	GPL1322	Drosophila melanogaster	fly head
GSM440101	0.267	0.050	GPL1322	Drosophila melanogaster	fly head
GSM101540	0.268	0.051	GPL72	Drosophila melanogaster	Heads of Drosophila received humidified air for 50 min followed by a single exp
GSM368570	0.268	0.051	GPL1322	Drosophila melanogaster	D.m. adult heads yw
GSM80135	0.268	0.051	GPL72	Drosophila melanogaster	adult head
GSM185064	0.270	0.053	GPL1322	Drosophila melanogaster	Fly Head
GSM266583	0.270	0.053	GPL1322	Drosophila melanogaster	Male fly heads, TORC neuronal resuce
GSM296980	0.270	0.053	GPL1322	Drosophila melanogaster	Diet:Untreated
GSM30828	0.270	0.053	GPL72	Drosophila melanogaster	whole head

Table 5: Human/mouse liver - zebrafish matches

	distance	pvalue	GPL	Species	Source
GSM280435	0.114	0.006	GPL1319	Danio rerio	Liver
GSM306849	0.118	0.006	GPL1319	Danio rerio	Dissociated transgenic zebrafish embryos [Tg(XlEef1a1:GFP)s854], flow sorted, 6 dpf
GSM280436	0.120	0.006	GPL1319	Danio rerio	Liver
GSM280438	0.122	0.007	GPL1319	Danio rerio	Liver
GSM280443	0.122	0.007	GPL1319	Danio rerio	Liver
GSM280437	0.127	0.007	GPL1319	Danio rerio	Liver
GSM220169	0.129	0.008	GPL1319	Danio rerio	Liver Female 25%carb
GSM306843	0.131	0.008	GPL1319	Danio rerio	Dissociated transgenic zebrafish embryos [Tg(XlEef1a1:GFP)s854], flow sorted, 4 dpf
GSM306851	0.133	0.008	GPL1319	Danio rerio	Dissociated transgenic zebrafish embryos [Tg(XlEef1a1:GFP)s854], flow sorted, 6 dpf
GSM280434	0.135	0.009	GPL1319	Danio rerio	Liver
GSM306847	0.135	0.009	GPL1319	Danio rerio	Dissociated transgenic zebrafish embryos [Tg(XlEef1a1:GFP)s854], flow sorted, 4 dpf
GSM280439	0.137	0.009	GPL1319	Danio rerio	Liver
GSM280441	0.137	0.009	GPL1319	Danio rerio	Liver
GSM306853	0.139	0.009	GPL1319	Danio rerio	Dissociated transgenic zebrafish embryos [Tg(XlEef1a1:GFP)s854], flow sorted, 6 dpf
GSM220175	0.143	0.010	GPL1319	Danio rerio	Liver Female 25%carb
GSM220174	0.150	0.011	GPL1319	Danio rerio	Liver Female 25%carb
GSM220182	0.150	0.011	GPL1319	Danio rerio	Liver Female 25%carb
GSM280442	0.150	0.011	GPL1319	Danio rerio	Liver
GSM280440	0.152	0.012	GPL1319	Danio rerio	Liver
GSM220171	0.160	0.013	GPL1319	Danio rerio	Liver Female 0%carb

Table 6: Human/mouse liver - rat matches

	distance	pvalue	GPL	Species	Source
GSM600627	0.068	0.002	GPL1355	Rattus norvegicus	Liver
GSM600629	0.070	0.002	GPL1355	Rattus norvegicus	Liver
GSM600649	0.070	0.002	GPL1355	Rattus norvegicus	Liver
GSM600625	0.074	0.003	GPL1355	Rattus norvegicus	Liver
GSM600651	0.074	0.003	GPL1355	Rattus norvegicus	Liver
GSM600630	0.076	0.003	GPL1355	Rattus norvegicus	Liver
GSM600771	0.076	0.003	GPL1355	Rattus norvegicus	Liver
GSM600574	0.078	0.003	GPL1355	Rattus norvegicus	Liver
GSM600591	0.078	0.003	GPL1355	Rattus norvegicus	Liver
GSM600624	0.080	0.003	GPL1355	Rattus norvegicus	Liver
GSM600721	0.080	0.003	GPL1355	Rattus norvegicus	Liver
GSM600840	0.080	0.003	GPL1355	Rattus norvegicus	Liver
GSM127076	0.082	0.003	GPL1355	Rattus norvegicus	Rat liver, left lateral lobe
GSM263074	0.082	0.003	GPL1355	Rattus norvegicus	2628 Propiconazole 500ppm Liver
GSM600628	0.082	0.003	GPL1355	Rattus norvegicus	Liver
GSM600631	0.082	0.003	GPL1355	Rattus norvegicus	Liver
GSM600775	0.082	0.003	GPL1355	Rattus norvegicus	Liver
GSM593239	0.084	0.003	GPL1355	Rattus norvegicus	Kinetic Study - at-RA 16 hr - Expt 2
GSM600647	0.084	0.003	GPL1355	Rattus norvegicus	Liver
GSM600720	0.084	0.003	GPL1355	Rattus norvegicus	Liver

Table 7: Human/mouse liver - C. elegans matches

	distance	pvalue	GPL	Species	Source
GSM28399	0.137	0.009	GPL200	Caenorhabditis elegans	Whole animal
GSM28395	0.141	0.009	GPL200	Caenorhabditis elegans	Whole animal
GSM28398	0.141	0.009	GPL200	Caenorhabditis elegans	Whole animal
GSM28407	0.141	0.009	GPL200	Caenorhabditis elegans	Whole animal
GSM28406	0.143	0.010	GPL200	Caenorhabditis elegans	Whole animal
GSM28402	0.144	0.010	GPL200	Caenorhabditis elegans	Whole animal
GSM28397	0.152	0.012	GPL200	Caenorhabditis elegans	Whole animal
GSM28405	0.154	0.012	GPL200	Caenorhabditis elegans	Whole animal
GSM756619	0.160	0.013	GPL200	Caenorhabditis elegans	whole worm L3 lysates
GSM756620	0.162	0.014	GPL200	Caenorhabditis elegans	whole worm L3 lysates
GSM756622	0.162	0.014	GPL200	Caenorhabditis elegans	whole worm L3 lysates
GSM756618	0.169	0.015	GPL200	Caenorhabditis elegans	whole worm L3 lysates
GSM756623	0.171	0.016	GPL200	Caenorhabditis elegans	whole worm L3 lysates
GSM756621	0.173	0.017	GPL200	Caenorhabditis elegans	whole worm L3 lysates
GSM378585	0.175	0.017	GPL200	Caenorhabditis elegans	Young adult (24 h post-vulval crescent L4 stage) whole nematodes were reared at
GSM562076	0.175	0.017	GPL200	Caenorhabditis elegans	Adult C. elegans exposed to 1 <ce><bc>g/ml IVM and 0.01% v/v DMSO for 4
GSM378583	0.181	0.019	GPL200	Caenorhabditis elegans	Young adult (24 h post-vulval crescent L4 stage) whole nematodes were reared at
GSM562081	0.181	0.019	GPL200	Caenorhabditis elegans	Adult C. elegans exposed to 0.01% v/v DMSO for 4 hrs
GSM28391	0.183	0.019	GPL200	Caenorhabditis elegans	Whole animal
GSM756617	0.186	0.021	GPL200	Caenorhabditis elegans	whole worm L3 lysates

Table 8: Human/mouse liver - Drosophila matches

	distance	pvalue	GPL	Species	Source
GSM419411	0.192	0.022	GPL1322	Drosophila melanogaster	Whole 7-8 old male
GSM419410	0.194	0.023	GPL1322	Drosophila melanogaster	Whole 7-8 old male
GSM419412	0.194	0.023	GPL1322	Drosophila melanogaster	Whole 7-8 old male
GSM216506	0.200	0.025	GPL1322	Drosophila melanogaster	whole bodies of 42 day old pyd male flies, replicate 2
GSM216513	0.200	0.025	GPL1322	Drosophila melanogaster	whole bodies of 42 day old crol male flies, replicate 1
GSM530095	0.200	0.025	GPL1322	Drosophila melanogaster	young flies fed with 100uM curcumin
GSM530094	0.203	0.027	GPL1322	Drosophila melanogaster	young flies without curcumin-feeding
GSM588948	0.203	0.027	GPL1322	Drosophila melanogaster	3 sets of 20 adults each was prepared from adult females from the correspondi
GSM530097	0.205	0.028	GPL1322	Drosophila melanogaster	old flies fed with 100uM curcumin
GSM216522	0.207	0.028	GPL1322	Drosophila melanogaster	whole bodies of 42 day old CG9238 male flies, replicate 1
GSM216515	0.209	0.029	GPL1322	Drosophila melanogaster	whole bodies of 42 day old crol male flies, replicate 2
GSM588947	0.209	0.029	GPL1322	Drosophila melanogaster	3 sets of 20 adults each was prepared from adult females from the correspondi
GSM216501	0.213	0.031	GPL1322	Drosophila melanogaster	whole bodies of 42 day old control male flies, replicate 1
GSM216505	0.213	0.031	GPL1322	Drosophila melanogaster	whole bodies of 42 day old pyd male flies, replicate 1
GSM419417	0.215	0.032	GPL1322	Drosophila melanogaster	Whole 7-8 old male
GSM676061	0.217	0.033	GPL72	Drosophila melanogaster	2-3 day old adults
GSM216523	0.219	0.034	GPL1322	Drosophila melanogaster	whole bodies of 42 day old CG9238 male flies, replicate 2
GSM216531	0.219	0.034	GPL1322	Drosophila melanogaster	whole bodies of 42 day old esg male flies, replicate 2
GSM588946	0.219	0.034	GPL1322	Drosophila melanogaster	3 sets of 20 adults each was prepared from adult females from the correspondi
GSM216530	0.221	0.035	GPL1322	Drosophila melanogaster	whole bodies of 42 day old esg male flies, replicate 1

2 Precision-recall of matching arrays

Next step is to quantify this - or show top arrays matching in each species or platform. For example, how many brain samples are there with p<0.02. To do this we need to collect the metadata across the full corpus. This will be done in a rough way by collecting *terms* or keywords related with specific tissues and seraching the source, chararcteristics and title for any of these *terms*. The top 7500 arrays for each tissue were manually re-curated to ensure that the *terms* set was appropriate. In addition, specific GSEs that had insufficient metadata to identify the source were individually examined. These were assigned on a case by case basis.

```
> brain.terms<-c(
+   "brain", "Amygdala", "Cortex","Lobe", "hippocamp",
+   "putamen", "hypothalamus", "Hypothlamus", "cerebellum", "Striatum",
+   "DLPFC", "cerebellum","medulla", "Gyrus",
+   "Glioblastoma", "Accumbens", "astrocytoma", "Medulloblastoma",
+   "oligodendrogliomas","cerebrum", "Ventral tegmental", "cerebrum",
+   "stria terminalis", "periaqueductal gray", "thalamus", "cerebellar",
+   "substantia_nigra", "Caudate", "ventral tegmental area",
+   "vestibular_nuclei_superior", "ventral_tegmental_area",
+   "Gloubus Pallidum", "Globus pallidus",
+   "subthalamic_nucleus", "corpus_callosum",
+   "Substantia Nigra", "spinal",
+   "nodose_nucleus", "corticotectal",
+   "Paraventricular", "cortical neurons",
+   "pons", "callosal", "oculomotor nucleus",
+   "Lateral geniculate nucleus",
+   "pituitary_gland", "trigeminal_ganglia",
+   "dorsal_root_ganglia", "pars tuberalis",
+   "neuroblastoma", "dorsal root ganglion",
+   "Supraoptic nucleus", "glioma",
+   "astrocytes", "pineal gland",
+   "3ARS02080774b_globus_pallidus"
+               )
> brain.GSE<-c("GSE9443","GSE11100", "GSE15222", "GSE9566", "GSE13041", "GSE2817",
+              "GSE19402", "GSE4206","GSE17617", "GSE13353", "GSE4623", "GSE2547",
+              "GSE19332", "GSE6614")
> kidney.terms<-c("kidney", "renal", "glomerulus", "renal", "Papillary", "RCC")
> kidney.GSE<-c("GSE7869", "GSE5243", "GSE13065", "GSE5243", "GSE7869", "GSE1563")
> liver.terms<-c("liver", "hepatocellular", "hepatocyte",
+                "Hepatic", "hepatocytes",
+                "s854 2dpf GFP+", "s854 3dpf GFP+",
+                "s854 4dpf GFP+", "s854 6dpf GFP+",
+                "cholangiocarcinoma", "HepaRG")
> liver.GSE<-c("GSE6632", "GSE10493", "GSE6903", "GSE9012", "GSE4285", "GSE4740",
```

```
+              "GSE1088", "GSE1089", "GSE14712")
> lung.terms<-c("lung", "bronchus", "NSCLC", "Bronchial")
> lung.GSE<-c("GSE11056", "GSE11809", "GSE18083", "GSE7670", "GSE6135", "GSE4512",
+             "GSE21581")
> muscle.terms<-c(
+   "skeletal muscle", "Proximal muscle", "vastus lateralis", "Biceps Brachii",
+   "vastus laterlis", "Quadricep", "skeletal_muscle", "Paravertebral",
+   "Deltoid", "Rectus femoris", "triceps", "superior rectus", "Vastus Lateralus",
+   "Gastrocnemius", "Flexor carpi", "tibialis cranialis", "tibialis anterior",
+   "calf muscle", "carpi radialis brevis", "gracillus", "tibialis anterior muscle",
+   "gracilis", "latissimus dorsi", "Soleus muscle", "muscle", "Tongue", "heart",
+   "diaphragm", "Intestine", "atrial", "ventricular",  "Cardiac", "Ventricle",
+   "Cardiomyocytes", "Anterior tibialis")
> muscle.GSE<-c("GSM397580", "GSM397583", "GSE1551", "GSE1764", "GSE80", "GSE12580",
+               "GSE21610", "GSE897", "GSE3307", "GSE6970")
> spleen.terms<-c("spleen", "splen", "lymph")
> spleen.GSE<-c("GSE24350", "GSE6980", "GSE24350", "GSE16059", "GSM541854")
> tissue.terms<-list(
+   brain.terms, kidney.terms, liver.terms,
+   lung.terms, muscle.terms, spleen.terms
+     )
> tissue.GSE<-list(brain.GSE, kidney.GSE, liver.GSE,
+                  lung.GSE, muscle.GSE, spleen.GSE)
> names(tissue.terms)<-tissues
> names(tissue.GSE)<-tissues


> tissue.distance<-vector("list", length(tissues))
> tissueDistance.meta<-lapply(tissues, function(x){
+   cbind(tissueDistance[[x]],
+           GEO.metadata.matrix[
+             match(rownames(tissueDistance[[x]]),
+             GEO.metadata.matrix$GSM)
+               ,])})
> names(tissueDistance.meta)<-tissues
> for (i in 1:length(tissues)){
+   tissue<-tissues[i]
+   print(tissue)
+   tissueDistance.meta[[tissue]]$valid<-(rowSums(
+   sapply(tissue.terms[[tissue]], function(x){
+     grepl(x, apply(
+             tissueDistance.meta[[tissue]][
+               ,c("Title", "Source", "Characteristics")],
+             1, function(y){
+                      paste(y, collapse = " ")
+                      }
```

```
+                   ), ignore.case = TRUE)
+       })
+   ) > 0)
+   tissueDistance.meta[[tissue]]$valid[
+       tissueDistance.meta[[tissue]]$GSE %in% tissue.GSE[[tissue]]
+       ]<-1
+   }
[1] "brain"
[1] "kidney"
[1] "liver"
[1] "lung"
[1] "skeletal muscle"
[1] "spleen"
> # write tables for supplementary data
> for (i in tissues){
+   write.table(tissueDistance.meta[[i]][
+     ,c("GSM", "GSE", "GPL", "Title", "valid", "distance")],
+               file = paste(i, "retrievalTable.txt", sep = ""),
+               quote = FALSE, sep = "\t")
+ }
```

We can use this to calculate the mean average precision and plot the data as a histogram of the distance overlayed with precision, or as percision-recall plots

```
> MAP<-sapply(tissues, function(x){
+ mean(
+   (cumsum(tissueDistance.meta[[x]]$valid) /
+      (1:length(tissueDistance.meta[[x]]$valid)))[
+     (tissueDistance.meta[[x]]$valid == 1)
+     ])})
> MAP
          brain          kidney           liver            lung skeletal muscle
      0.6482726       0.4196622       0.7975898       0.4071706       0.5982356
         spleen
      0.1499265


> precision.recall.tissues<-lapply(tissues, function(x){
+   data.frame(
+     recall = (
+       cumsum(tissueDistance.meta[[x]]$valid) /
+         sum(tissueDistance.meta[[x]]$valid)
+         ),
+     precision = (
+       cumsum(tissueDistance.meta[[x]]$valid) /
+         (1:length(tissueDistance.meta[[x]]$valid))
```

```
> par(mfcol = c(2,3))
> for (i in 1:length(tissues)){
+ hist(tissueDistance[[i]][,"distance"],
+     breaks = seq(0,1,(sum(abs(2*tissueConsensus[,i])))^(-1)),
+     xlim = c(0,1), main = tissues[[i]],
+      xlab = "Scaled distance", ylab = "")
+ par(new = TRUE)
+ plot(tissueDistance.meta[[i]]$distance,
+      (cumsum(tissueDistance.meta[[i]]$valid) /
+         (1:length(tissueDistance.meta[[i]]$valid))),
+      ylim = c(0,1), xlim = c(0,1),
+      xaxt = "n", yaxt = "n", xlab = "",
+      ylab = "", type = "l", col = "red")
+ axis(4)
+ }
```

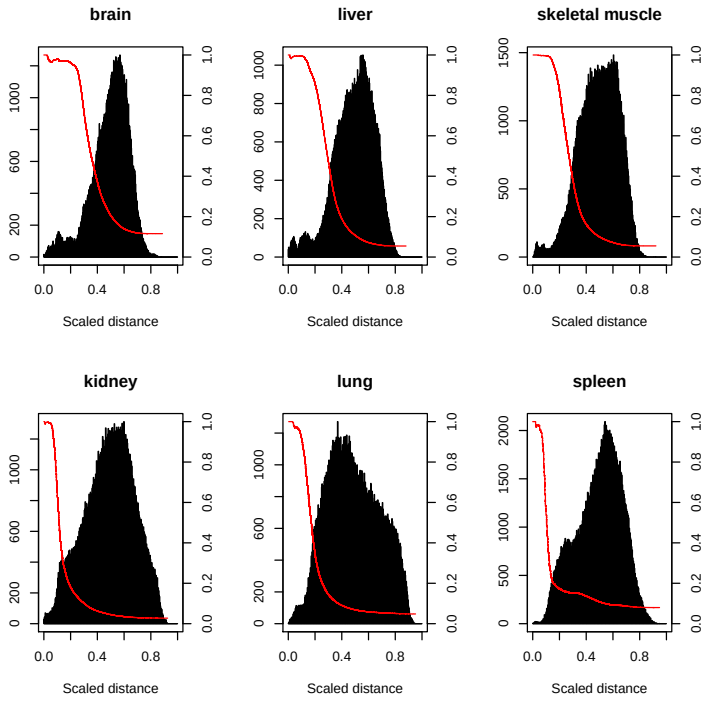

Figure 3: Frequency histograms of this distribution of the distance of the GEO corpus from each tissue consensus fingerprint (black, left axis), and the associtated retrieval precision (red, right axis)

```
> hist(tissueDistance[[1]][,"distance"],
+     breaks = seq(0,1,(sum(abs(2*tissueConsensus[,1])))^(-1)),
+     xlim = c(0,1), main = tissues[[1]],
+      xlab = "Scaled distance", ylab = "")
> par(new = TRUE)
> plot(tissueDistance.meta[[1]]$distance,
+      (cumsum(tissueDistance.meta[[1]]$valid) /
+         (1:length(tissueDistance.meta[[1]]$valid))),
+      ylim = c(0,1), xlim = c(0,1),
+      xaxt = "n", yaxt = "n", xlab = "",
+      ylab = "", type = "l", col = "red")
> axis(4)
```

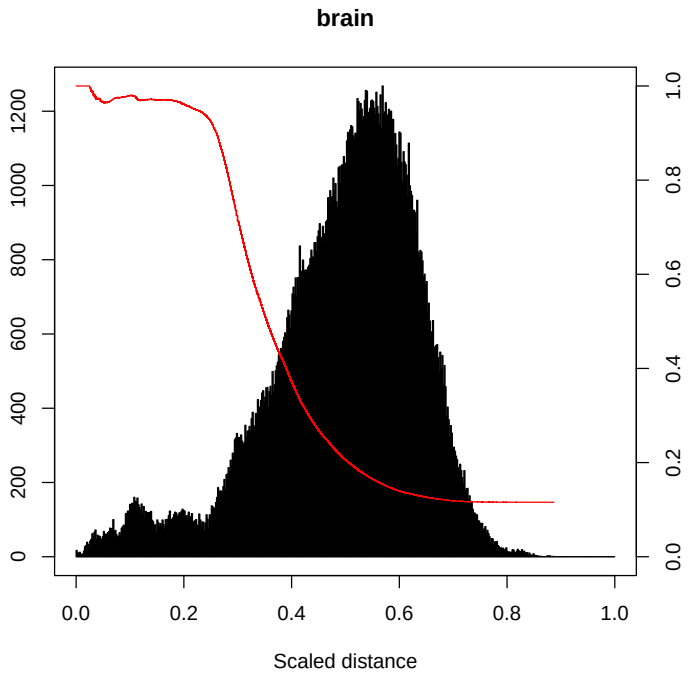

Figure 4: Frequency histograms of this distribution of the distance of the GEO corpus from the brain consensus fingerprint (black, left axis), and the associtated retrieval precision (red, right axis)

```
+         )
+             )
+   })
> names(precision.recall.tissues)<-tissues
> precision.recall.tissues.std<-lapply(
+   precision.recall.tissues, function(z){
+     approx(
+       x = z$recall,
+       y = z$precision,
+       xout = seq(0.01, 1, 0.01),
+       rule = 2
+     )$y
+   })
```

Finally the results can be summarized as a table

```
> # create table at 95% and 90% precision
> top95.table<-as.data.frame(matrix(nrow = length(tissues), ncol = 4))
> rownames(top95.table)<-tissues
> colnames(top95.table)<-c("Seed arrays", "Correct retrievals",
+                          "Platforms", "Species")
> top95.table$"Seed arrays"<-table(tissue.meta$Tissue)[tissues]
> for (i in tissues){
+   top95<-precision.recall.tissues[[i]]$precision > 0.95
+   top95.correct<-(top95 & (tissueDistance.meta[[i]]$valid ==1))
+   top95.table[i,"Correct retrievals"]<-sum(top95.correct)
+   top95.table[i,"Platforms"]<-length(
+     table(tissueDistance.meta[[i]]$GPL[top95.correct])
+     )
+   top95.table[i,"Species"]<-length(
+     table(tissueDistance.meta[[i]]$Species[top95.correct])
+     )
+ }
> top90.table<-as.data.frame(matrix(nrow = length(tissues), ncol = 4))
> rownames(top90.table)<-tissues
> colnames(top90.table)<-c("Seed arrays", "Correct retrievals",
+                          "Platforms", "Species")
> top90.table$"Seed arrays"<-table(tissue.meta$Tissue)[tissues]
> for (i in tissues){
+   top90<-precision.recall.tissues[[i]]$precision > 0.9
+   top90.correct<-(top90 & (tissueDistance.meta[[i]]$valid ==1))
+   top90.table[i,"Correct retrievals"]<-sum(top90.correct)
+   top90.table[i,"Platforms"]<-length(
+     table(tissueDistance.meta[[i]]$GPL[top90.correct])
+     )
+   top90.table[i,"Species"]<-length(
```

```
> plot.new()
> axis(1, seq(0,1,0.2))
> axis(2, seq(0,1,0.2))
> title(xlab = "recall", ylab = "precision")
> for (i in 1:length(tissues)){
+ tissue<-tissues[i]
+ lines(x = seq(0.01, 1, 0.01),
+        y = precision.recall.tissues.std[[tissue]],
+        col = rainbow(6)[i])
+ }
> legend(x = "topright", tissues,
+          text.col = rainbow(6)[1:6], bty= "n", cex = 0.75)
```

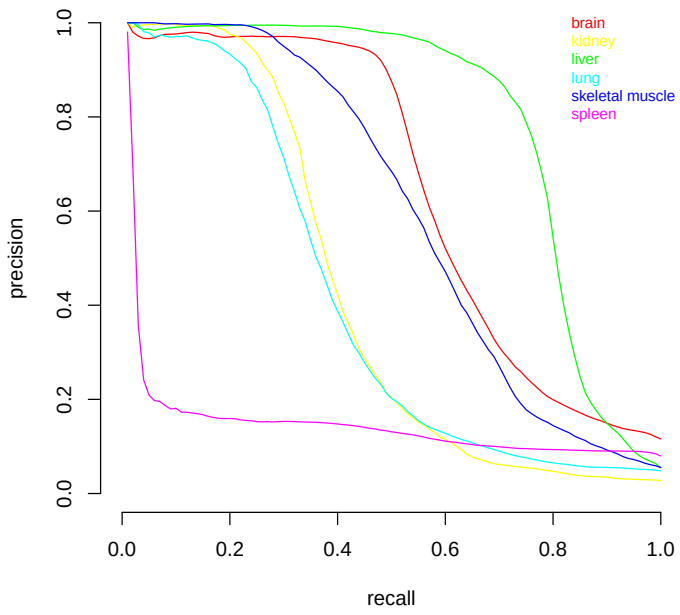

Figure 5: Precision-recall curves for matched tissue retrieval from the GEO database

```
+     table(tissueDistance.meta[[i]]$Species[top90.correct])
+     )
+ }
```

	Seed arrays	Correct retrievals	Platforms	Species
brain	50	9379	27	4
kidney	81	1227	16	3
liver	196	6032	26	4
lung	142	1621	14	3
skeletal muscle	29	3135	20	2
spleen	33	190	6	2

Table 9: Tissue retrieval 95pct precision

	Seed arrays	Correct retrievals	Platforms	Species
brain	50	10693	29	5
kidney	81	1384	16	3
liver	196	6945	27	4
lung	142	2092	17	3
skeletal muscle	29	3731	23	3
spleen	33	227	9	2

Table 10: Tissue retrieval 90pct precision